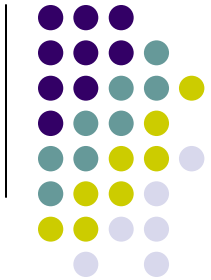
USIIS
Utah Statewide Immunization Information System
Record Matching Project

Nancy McConnell
Chris Pratt

Agenda

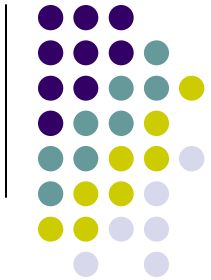


- USIIS Background
- USIIS Features

USIIS Record Matching Project

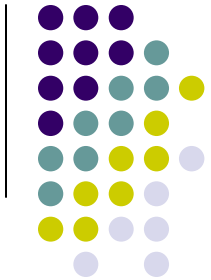
- Problem
- Solution: Planning, Design Process, Evaluation
- Description of Match Algorithm
- Solution: Deployment
- Results
- Retrospect
- Looking Ahead

USIIS Background
Utah's Immunization Registry



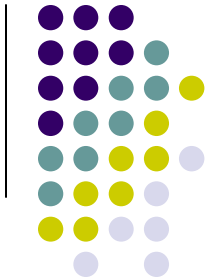
- Operational milestones
  - 1995: CDC delivered USIIS to Utah a pilot study component
    - Limited deployment to a subset of local health departments
  - 1998: UDOH released re-engineered USIIS
    - Overhaul to database, application, matching algorithm
    - Deployed to local health departments, a few private providers
  - 1999: Web application deployed
    - Private provider recruitment
  - 2000-2001: Intermountain Healthcare commitment to USIIS and introduction into clinics
    - Demonstrated time savings
    - Demonstrated improved immunization rates
  - 2003: Emergency Incident Management System
  - 2004: Adult data rule took effect
  - 2007: Released new record matching algorithm and loading program

USIIS Background, continued



- Users
  - Over 670 organizations
    - Private providers
    - Utah Local Health Departments
    - Federally qualified health centers, rural health centers, community health centers, other public clinics
    - Schools and daycares
  - Three active Users Groups
    - Bear River Health District
    - Northern Utah – Ogden, Davis
    - Salt Lake

USIIS Background, continued

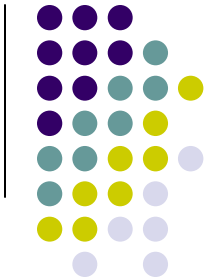


- Methods of data entry
  - Web application data entry
  - Data interfaces from providers' systems
    - Flat file (proprietary format)
    - HL7 (2001)
  - Routine UDOH data loads
    - Vital Records (1998)
    - WIC (1998)
  - Real-time integration with Intermountain Healthcare HELP and ADT systems (2002)
    - Obtain daily patient index record, as patients register for any service
- Some data characteristics
  - Patient records: 2,656,665
  - Vaccination records: 17,272,989
  - All children born in Utah since 1998
  - Utah children $<$ 6 years of age with $\geq$ 2 vaccinations: 59.8 %

USIIS Features

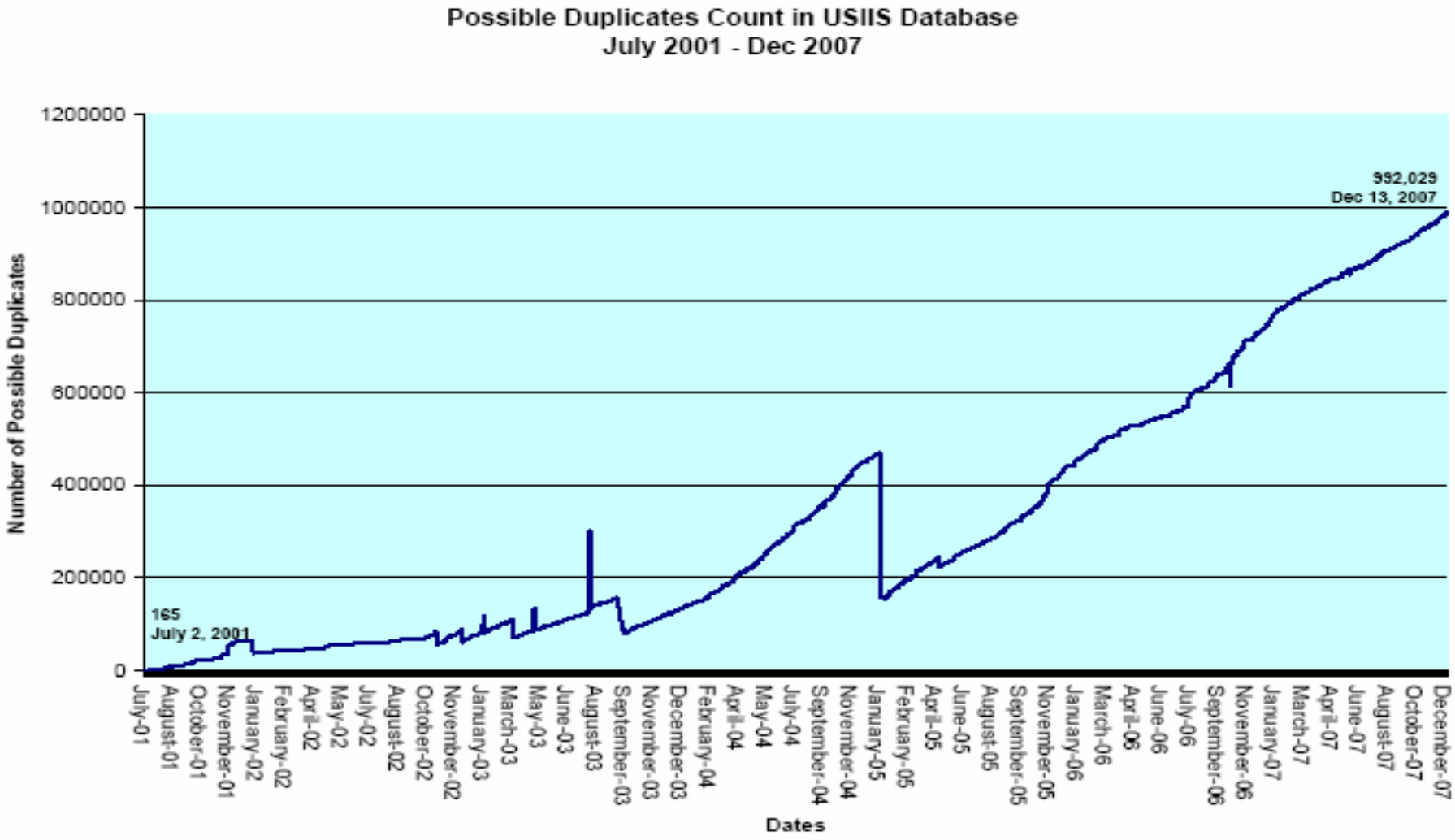
- Daily clinical decisions made based on USIIS information
- Key features
  - Immunization histories for Utah citizens—consolidated across providers
  - Immunization forecast
    - Immunizations due or past due based on CDC schedules
  - Patient reports
    - School Immunization record
    - Personal Immunization record and history
  - Clinic reports
    - Doses administered
    - Batch Forecast and Reminder-Recall
  - Vaccine inventory management
- UDOH data use
  - Immunization coverage rates
  - Immunization documentation required by federal programs
  - HEDIS measures

Problem
Duplicate and "Possible Duplicate" Records

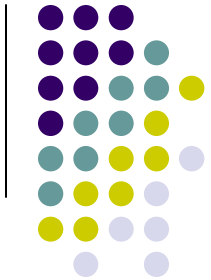


- Problem
  - Number of duplicate records: 20.9% of USIIS records
  - Number of "possible duplicate" records: 27% of USIIS records
- Impact of Problem
  - Difficult to find patients in USIIS—key barrier to provider participation
  - Risk of over-immunization—unable to find reliable patient record
    - Cost of unnecessary immunizations
    - Risk of adverse effects on patients
  - Immunization coverage rates understate actual levels
- Source of Problem
  - Patients submitted with minimal an variable data
  - Same patient entered with different names
    - Aliases, nicknames
    - Cultural variations
    - Family changes—e.g., divorce, adoption, multiple births
  - Information entered with typos—names, birth dates, etc.

Problem
Growth of “Possible Duplicate” Records

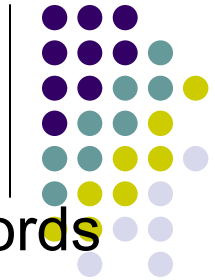


Problem
Internal Record Management



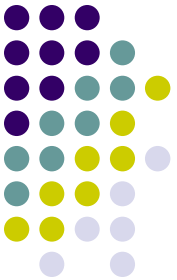
- Manual review of “possible duplicate” records
  - Unmanageable—operated in react mode
  - Costly
- High volume of Help Desk tickets to find or merge duplicate records
  - Customer dissatisfaction
  - Costly
- Poor record audit trail
  - Limited troubleshooting capability
  - Precluded data quality initiatives

Solution: Planning



- Reduce quantity of duplicate and possible duplicate records
  - Design and deploy a better record match algorithm
- Goals for new Match Algorithm and record processing
  - Record processing: Batch and real-time
  - Performance: 40-200 records per minute
  - Reduce the number of possible duplicates by 30% - 75%
  - Reduce duplication rate to 2% - 8%
  - Design
    - Built-in record auditing
    - Modular
    - Parameter-driven algorithm for quick and easy adjustment
- Project approach
  - Development assignments across entire programming team
  - Testing
    - Unit, integrated and system testing
    - Documented test cases and expected results

Solution: Design Process



- Developed two “competing” record match algorithms
  - Deterministic and probabilistic components in both
  - #1 focus on human record matching behaviors
  - #2 focus on statistical analysis with rule exceptions
- Developed “Gold Standard”
  - Standard by which competing algorithms would be compared
  - Based on Vital Records
  - Enhanced by human expert knowledge of USIIS data (manual)
- Evaluated algorithms with two data sets
  - Ran records through both algorithms and Gold Standard
  - Evaluated results
    - Compared results to Gold Standard results for record matching accuracy
    - Measured Sensitivity (90%) and Specificity (100%)—to CDC guidelines
  - Results indicated significant improvement in record matching
  - Selected Match Algorithm #1 to implement

Solution: Design Process
Algorithm Evaluation

- Match Algorithm – record categories:
  - Merge (match – merge with an existing record)
  - Possible duplicate (possible match - manual review)
  - Insert (no match – insert as a unique record).
- Gold Standard – two possible results:
  - Merge/match
  - Insert/non-match.

Incoming Records Matching Results

		MM	MP	MI
(True Positive) TP =	105	105	9	0
(True Negative) TN =	54	II	IP	IM
(False Positive) FP =	0	30	24	0
(False Negative) FN =	9			
	168	135	33	0

	Gold VR	USIIS
MM =	Merge	Merge
MP =	Merge	Possible
MI =	Merge	Insert
II =	Insert	Insert
IP =	Insert	Possible
IM =	Insert	Merge

USIIS	Gold VR Match	Gold VR Non.Match	
Match	a 105 True+	b 0 False+	105
Non-match	c 9 False -	d 54 True -	63
	114	54	168

TP / TP + FN	Sensitivity	92%
TN / TN + FP	Specificity	100%
TP+TN/TP+FP+TN+FN	Accuracy	95%
TP / TP + FP	Precision	100%
FP / TN + FP	False Positive	0%

- Outcomes from the Gold Standard and Match Algorithm were formatted into a 2 x 2 table in order to calculate evaluation metrics.
- Possible duplicates (MP & IP) were included as non-match for the Match Algorithm.

Solution: Design Process
Algorithm Evaluation

- Consulted with University of Utah Center of Excellence in Public Health Informatics (COE)
  - COE reviewed and supported the USIIS team's approach and results
  - COE conducted independent study using USIIS data
  - COE recommended some tweaks to the Match Algorithm
- Documented Match Algorithm specifications

Load Matching Queries
PROBABILISTIC BLOCKING

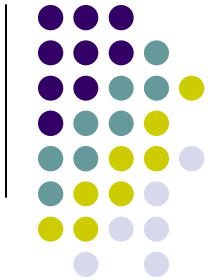
against pat_alias table

				Key Identifiers						Person Identifiers		
Order	query code	Execute condition	Identifer types	Usiis id	prov id	pat id	end date	ssn	medicaid id	fname	mname or minit	minit
1	P0	usiis_id not null and > 0		x								
2	P1				x	x						
3	P2	ssn not null						x				
4	P3	medicaid # not null							x			
5	P4	middle not null								x	x	
6	P5									x		
7	P6	middle not null								first 3	x	
7		middle not null									x	
7		middle not null								x	x	
8	P7									first 3		
9	P8											
10	P9	middle not null										
11	P10									first 3		
11										x		
11												
12	P11	phone not null								x		
12		phone not null										
12		phone not null										
13	P12	street not null								x		
13		street not null										
13		street not null										
14	P13	maiden <> UNKNOWN								x		
14		maiden <> UNKNOWN										
14		maiden <> UNKNOWN										

against purge_pat table

				Key Identifiers	Person Identifiers

Match Algorithm Description
USIIS Data Characteristics



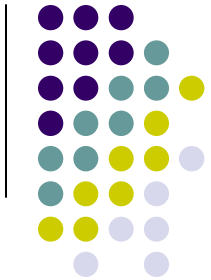
- How often data elements are present in USIIS records
  - Use valid values populated ≥ 20% of the time
  - Use valid values populated <20% of the time—if closely linked to another data element that is populated

  ☑ Match Algorithm based on USIIS-specific data characteristics
- How data elements relate in the real world, or, how users identify patients—in order of practice
  - Patient information—i.e., name, birth date, SSN
  - Parent information, especially the mother's
  - Other information that boosts confidence—e.g., gender, race, vaccination patterns

  ☑ Match Algorithm mimics human matching behavior via data groups
- How data elements may change
  - Transient data include address, phone, father and guardian
  - Non-matches carry no penalty; matches indicate strong link

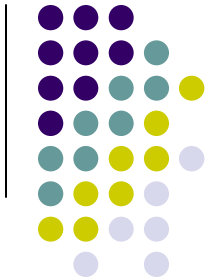
  ☑ Match Algorithm applies weights to data groups

Match Algorithm Description
Deterministic Match



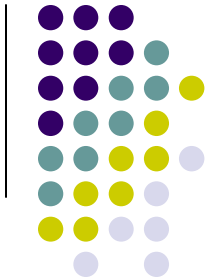
- Deterministic match: If exact match found, stop
- Compare data elements for matches
  - First name, last name, date of birth and USIIS ID
  - First name, last name, date of birth and SSN
  - First name, last name, date of birth and middle name
- If exact match not found, conduct Probabilistic match

Match Algorithm Description
Probabilistic Match - Data fields



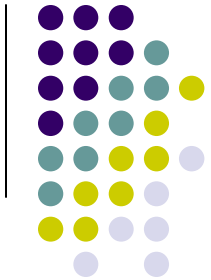
- Series of queries comparing data elements
- Compare data elements for matches
  - Key identifiers
    - USIIS ID
    - Provider ID and Provider Patient ID
    - SSN
    - Medicaid ID
  - Person identifiers
    - Patient first and middle names, suffix
    - Birth date
  - Family identifiers
    - Patient last name
    - Mother first, middle, last and maiden names

Match Algorithm Description
Probabilistic Match - Data fields, continued



- Data element groups, continued
  - Transient identifiers
    - Father and guardian first, middle and last names
    - Street address, city, state, zip and phone
  - Booster identifiers
    - Gender
    - Ethnicity
    - Race
    - Rare first or middle names
    - Number of matching vaccines
    - Vaccine record overdose
- Data group weights—developed considering
  - Human record matching behavior
  - USIIS data characteristics
  - Cultural patterns within USIIS data

Match Algorithm Description
Probabilistic Match - Special Conditions

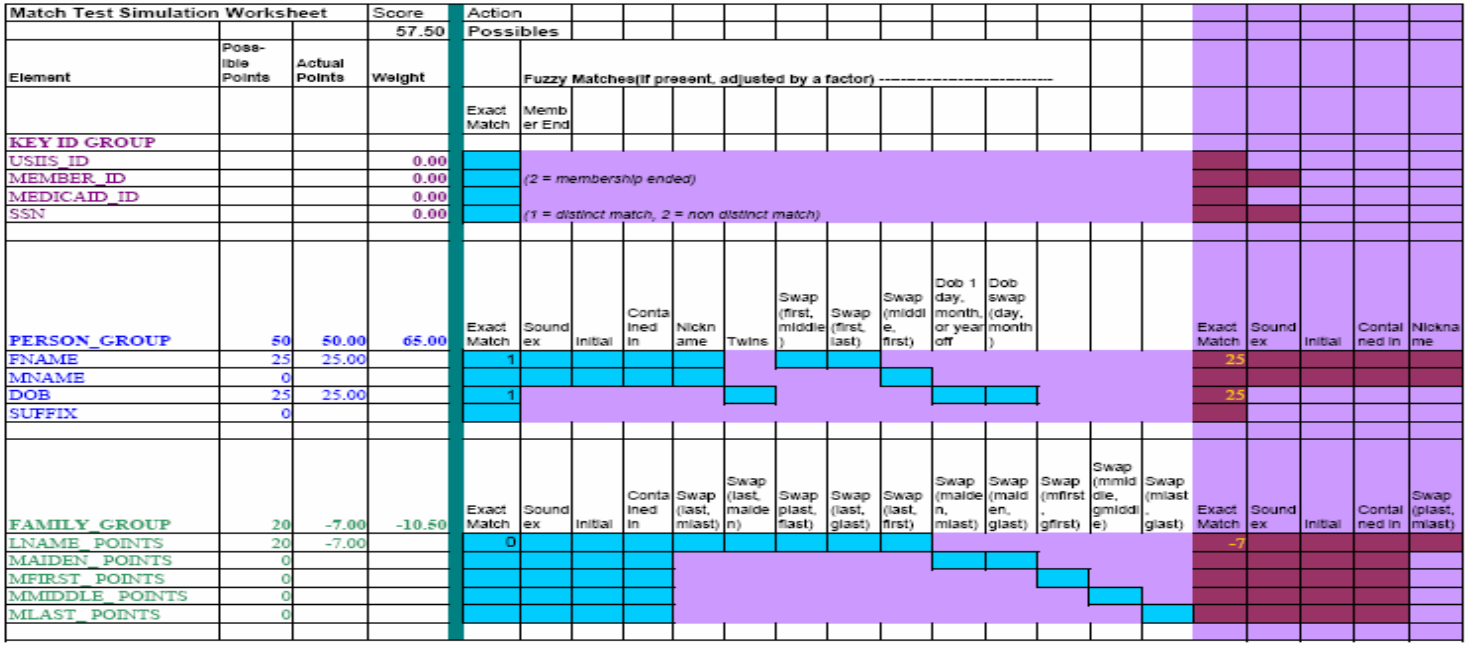


- Also, compare data elements for special conditions
  - “Contained in”—partial text
    - Name fields: Patient, mother maiden, father, guardian
  - Common nicknames
    - Patient first and middle names
  - Swapped names
    - Patient first name with: Patient last, patient middle
    - Patient last name with: Mother last, maiden, father last, guardian last
    - Father first name with: Father last, patient last
    - Guardian first & middle names with: Mother and father first & middle
    - Guardian last name with: Mother last, maiden, father last
  - Soundex and Initial in name
    - Name fields: Patient, mother, maiden, father, guardian
  - Birth date
    - Typos: One day off, one month off, or one year off
    - Swapped fields: Day and month
    - Multiple births: Greater reliance on patient name for a match

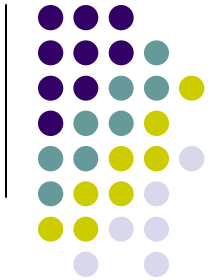
Match Algorithm Description
Scoring Schema

- Scoring schema
  - Data groups, representing USIIS data characteristics
  - Points
    - Applied to each data element in each data group
    - Positive or negative (penalty)
  - Weights: Applied to each data group
  - Total match score =

$$\sum_{\text{data groups}} \{\text{data group weight} * (\sum_{\text{data elements}} \text{data element point})\}$$

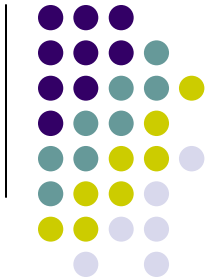


Solution: Putting It All Together



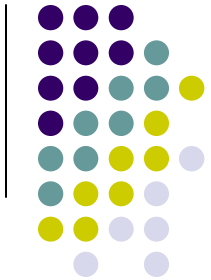
- Incoming patient record is compared to existing USIIS records using the Match Algorithm
- Match Algorithm assigns scores to record comparisons
- Record status is determined
  - New record: Match score < 49
    - Insert new record
  - Existing record: Match score ≥ 86
    - Update existing record
  - Possible existing record: 49 ≥ Match score ≤ 85
    - Store as possible duplicate for manual review

Solution: Deployment



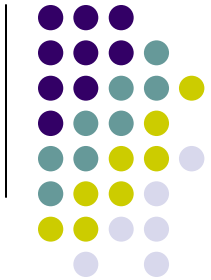
- Released December 15, 2007
  - All incoming records began running through the Match Algorithm
    - Web application
    - Interface data files
- Reloaded Vital Records (1998-current)
- Ran existing possible duplicate records through the Match Algorithm
  - Deleted “useless” possible duplicate records, according to business rules
    - Old vital records
    - WIC records without immunizations & whose patient age was outside eligible service range

Solution: Deployment, continued



- Emergence of new problems
  - Record auditing features of new record loading program revealed data quality issues previously hidden
    - “Orphan vaccinations”
    - Vaccinations administered before patient’s date of birth
    - Vaccinations administered in the future
    - Duplicate data within providers’ interface files
  - Modified loading program to better manage data quality
    - Improved data validation and standardization
  - Working with providers and vendors to improve quality of interface data files

Results

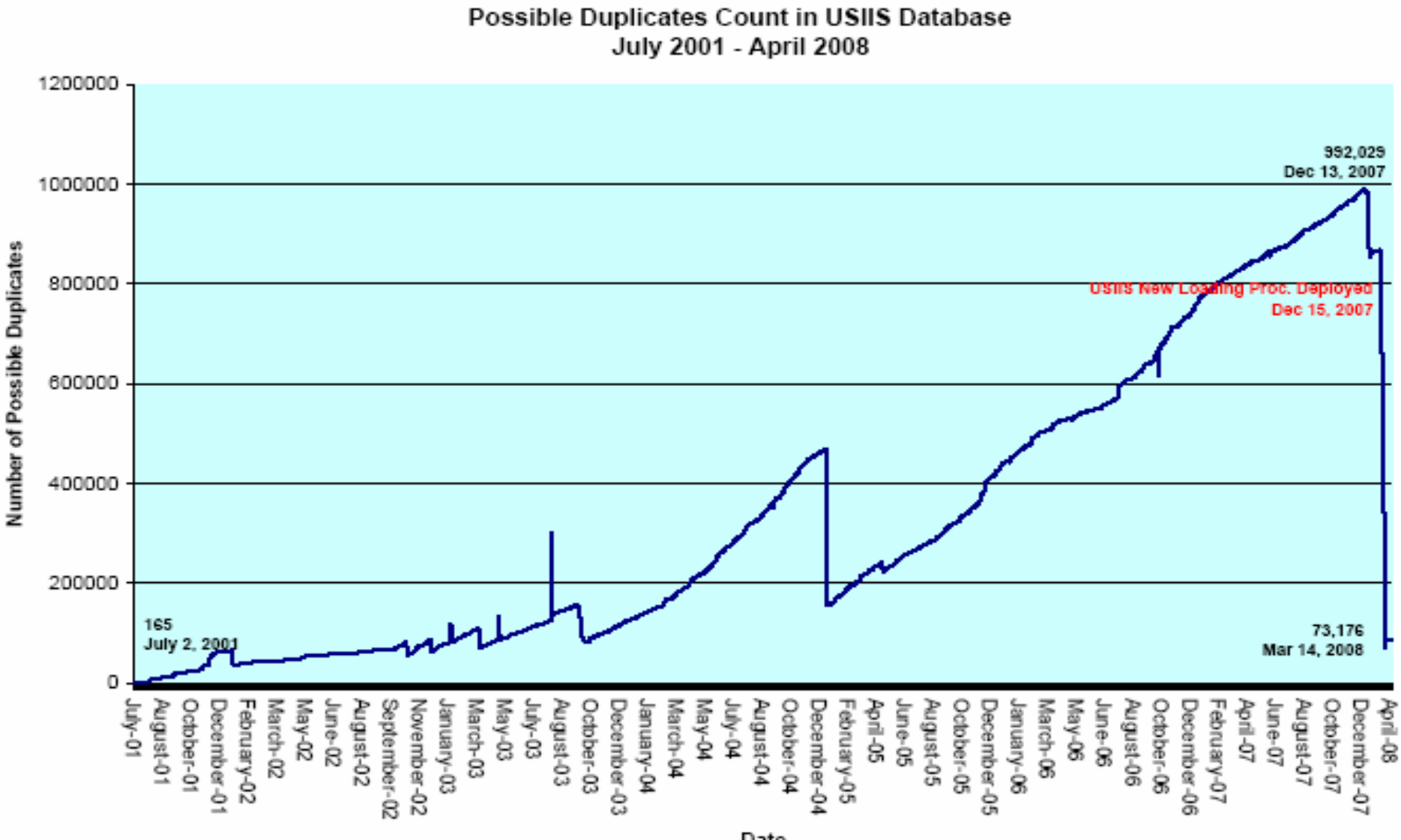


- Measurable goals

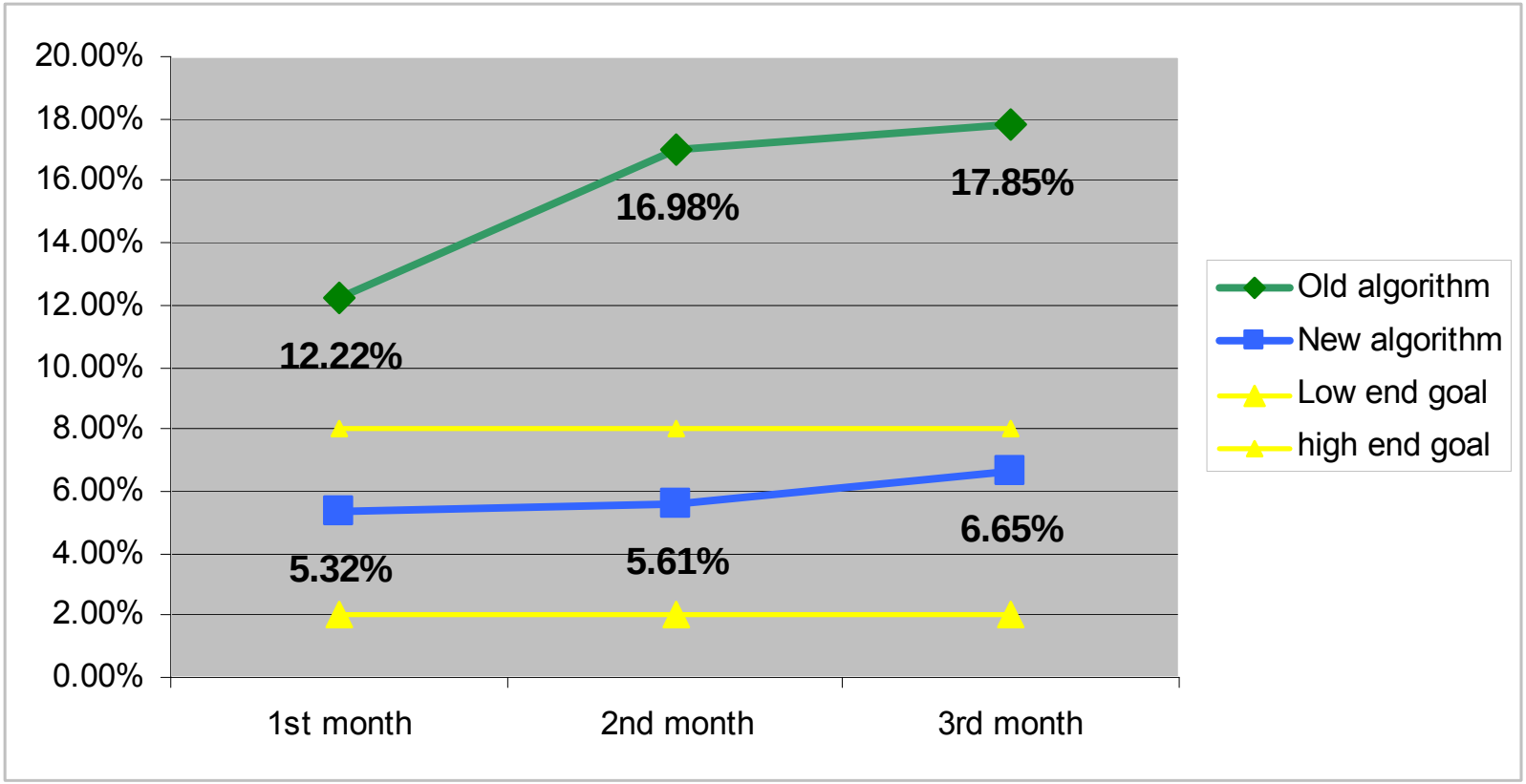
Measure	Units	Pre Value	Goal Range	Post Value	Results	Summary
Performance	Records per Minute	411	40 - 200	178	58% decrease	Close to high-end goal.
Possible Duplicates	Quantity	992,029	30-75% reduction	73,176	92.6% reduction	Exceeded high-end goal by 23%.
Possible Duplicate Rate	Records per load	18.10%	Unspecified	3.37%	81% reduction	Significant decrease.
Duplicate Rate	Last 3 month period	15.68%	2% - 8%	5.86%	63% reduction	Attained mid-range of goal.

- Auditing features
  - In use for troubleshooting
  - In use to provide data quality information to data-submitting vendors

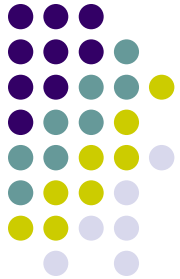
Results
Possible Duplicates



Results
Duplicate Rate

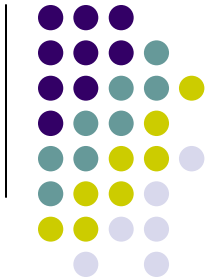


Retrospect



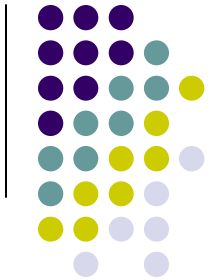
- USIIS is a software product for making clinical decisions
  - Product must be accurate and meet customer needs
  - Product must be thoroughly tested prior to release
- Algorithm
  - Deterministic-Probabilistic algorithm was a good solution
  - Developed a novel approach to record matching
    - Programming expert human logic
    - Based on USIIS data characteristics
- Project
  - Programming assignments
    - Learning curve
    - Payoff
  - Test cases
    - Some resistance
    - Payoff

Looking Ahead



- Match Algorithm improvement
  - Iterative improvement
    - Address “left over” possible duplicate records
- University of Utah COE
  - Continue working relationship
- Publicize and share
  - Share information about the algorithm and its design
- Data quality initiatives
  - Use record auditing capability to improve quality of submitted data
- Implement USIIS enhancements previously on hold
  - Making previously protected fields editable
  - Gradually enabling users the ability to resolve possible duplicate records

Credits



- USIIS Team!
  - JC Alexander
  - Eric Anderson
  - Marie LeFevre
  - Zhiwei Liu
  - Keyi Niu
  - Tom Romney
  - Sandy Schulthies
  - Yukiko Yoneoka
- Champions and sponsors!
  - Wu Xu
  - Barry Nangle
  - Larry Cook, University of Utah
  - Nancy Pare
  - George Delavan